BreastCancer_Analysis

Hrithik Kumar

2023-11-19

Introduction

In this report, we will discuss the Breast Cancer data set present in the *mlbench* package. The data set contains 699 rows and 11 columns. There are a few NA values as well in the data set, for simplicity, we will omit these values. The data set contains information about characteristics of the Breast tissue sample collected using Fine Needle Aspiration Cytology (FNAC). These characteristics, such as uniformity of cell and size of cell, are measured on a scale of 1-10. These variables are stored as factors, we will convert them to numeric variables for our analysis. Finally, we have a column named *Class* which denotes whether the tissue sample is Benign or Malignant. We aim to create a classifier which can predict whether a tissue sample is benign or malignant.

```
##   Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 1            5         1          1             1            2           1
## 2            5         4          4             5            7          10
## 3            3         1          1             1            2           2
## 4            6         8          8             1            3           4
## 5            4         1          1             3            2           1
## 6            8        10         10             8            7          10
##   Bl.cromatin Normal.nucleoli Mitoses Class
## 1           3               1       1     0
## 2           3               2       1     0
## 3           3               1       1     0
## 4           3               7       1     0
## 5           3               1       1     0
## 6           9               7       1     1
```

Exploratory Data Analysis

First, let's start with understanding the data set. After removing the NA values, the data set contains 683 rows and 10 columns. To begin our analysis, we will have a look at the mean and variance of the nine variables representing each tissue sample.

```
## Number of Benign and Malignant tissue samples. (0: Benign, 1: Malignant)
```

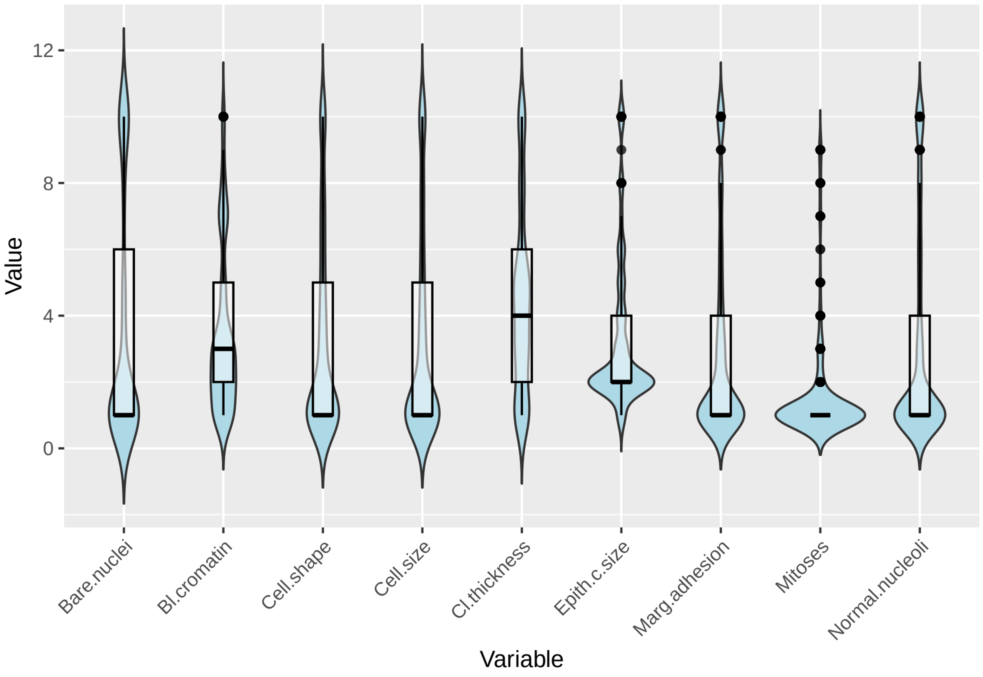
```
##
##   0   1
## 444 239
```

```
## Mean values of Each Variable:
```

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##        4.442167        3.150805        3.215227        2.830161        3.234261
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses
##        3.544656        3.445095        2.869693        1.582723
## Variance of Each Variable:
```

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##        7.956694        9.395113        8.931615        8.205717        4.942109
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses
##       13.277695        6.001013        9.318772        2.677531
```

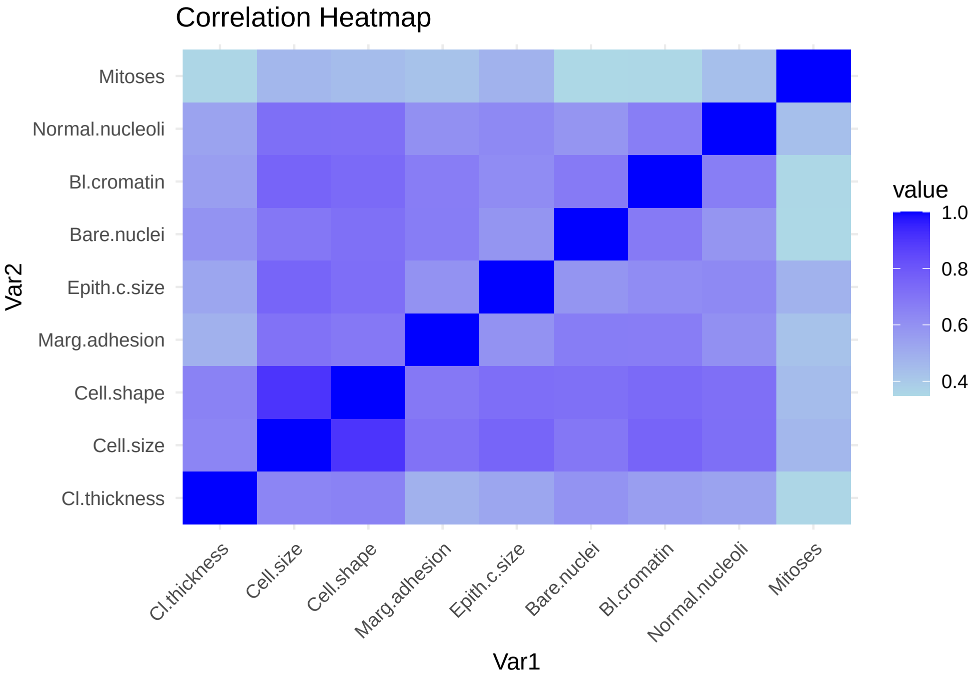
The Mean and variance do give us a general idea about the distribution of each variable, however, it does not provide us the whole picture. Mean and variance could be heavily influenced by outliers, which could lead to incorrect assumptions about the population based on the sample data. To better understand the distribution of each of these variables, a box and violin plot will be more helpful.



The above plot highlights very crucial information about the data set. Firstly, Distribution of “Mitoses” is heavily skewed and consists of many outliers. Similarly, Epithelial cell size, Normal nucleoli and Marginal adhesion have skewed distributions, but they are slightly less skewed and contain less number of outliers compared to Mitoses. The thick bottom end of the violins indicates that majority of these values lie around the median. Cell shape and Cell size have very similar distributions. Both the variables have a minimum value of 1, which is also the median value of the variables. Moreover, the 3rd Quantile of both the variables is also same, which is 5. Next, let’s have a look at the covariance matrix to get an idea about how these variables are related to each other.

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## Cl.thickness        7.956694  5.554922   5.508800      3.941776     3.283363
## Cell.size           5.554922  9.395113   8.310604      6.207468     5.134708
## Cell.shape          5.508800  8.310604   8.931615      5.872385     4.799947
## Marg.adhesion       3.941776  6.207468   5.872385      8.205717     3.786179
## Epith.c.size        3.283363  5.134708   4.799947      3.786179     4.942109
## Bare.nuclei         6.096061  7.725660   7.774099      7.000264     4.744656
## Bl.cromatin         3.826365  5.673248   5.383535      4.691541     3.366253
## Normal.nucleoli     4.598758  6.730824   6.550081      5.274024     4.268107
## Mitoses             1.636389  2.334281   2.185249      1.992076     1.750388
##                 Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
## Cl.thickness       6.096061    3.826365        4.598758 1.636389
## Cell.size          7.725660    5.673248        6.730824 2.334281
## Cell.shape         7.774099    5.383535        6.550081 2.185249
## Marg.adhesion      7.000264    4.691541        5.274024 1.992076
## Epith.c.size       4.744656    3.366253        4.268107 1.750388
## Bare.nuclei       13.277695    6.075403        6.499229 2.080978
## Bl.cromatin        6.075403    6.001013        4.977439 1.417672
## Normal.nucleoli    6.499229    4.977439        9.318772 2.183083
## Mitoses            2.080978    1.417672        2.183083 2.677531
```

The Covariance matrix consists of only positive values, indicating a positive relationship among all the variables. Cell size and Bare nuclei show strong correlation, similarly, Marginal adhesion and Cell shape also show strong correlation. These values are quite helpful, however, a heatmap would be more easy to interpret and extract insight from.



The above heat map is highlighting some very important relationships among the variables. The most obvious and intuitive result is the strong correlation between Cell size and Cell shape. Apart from this, we can confirm our previous interpretation of relation between Cell size and Marginal adhesion, and Cell size and Bara nuclei. Epithelial cell size, Bland cromatin and Normal nucleoli seem to have much stronger correlation with Cell size. On the other hand, Mitoses has very weak correlation with all the other variables. Similarly, Clump thickness also does not have very strong relation with any of the other variables, apart from Cell shape.

These insights point towards the possibility of only some of the variables contributing towards the target variable. We will calculate Generalized variance and Total variance of the data set to further understand the variability in the data.

```
## Generalised Variance:  47432.09
```

```
## 
## Total Variance:  70.70626
```

We have a large value of Generalized Variance suggesting a significant spread or variability in the data along specific directions. On the other hand, the Total Variance is quite small compared to Generalized variance, indicating that the data is quite concentrated or have less variability when all possible directions are considered. This means that there are strong trends or patterns in the data set that can be captured by a few subset of variables, while other variables contribute quite less to overall variability. We will now build classifiers based on our findings for the Class - benign or malignant.

Data Standardization

Before we implement models on the data set, we will first standardize the data set. We standardize the data set to make sure that variables with larger variances don't dominate the results. Subset selection and Regularization techniques work best on standardized data. After scaling the data, the mean of each variable should be 0, and variance should be 1. Moreover, the Covariance matrix of the scaled data should be equal to the Correlation matrix of the original data.

```
##
##  Variance of each Column:
```

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##               1               1               1               1               1
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses
##               1               1               1               1
```

```
##
##  Mean of each Column:
```

```
##    Cl.thickness       Cell.size      Cell.shape   Marg.adhesion    Epith.c.size
##    1.118949e-16    1.155667e-17   -5.288287e-17    6.517103e-17    8.209696e-18
##     Bare.nuclei     Bl.cromatin Normal.nucleoli         Mitoses
##   -1.786156e-17    3.187205e-17    3.671659e-17    2.566988e-17
```

```
##
##  Comparison of Covariance matrix and Correlation Matrix:
```

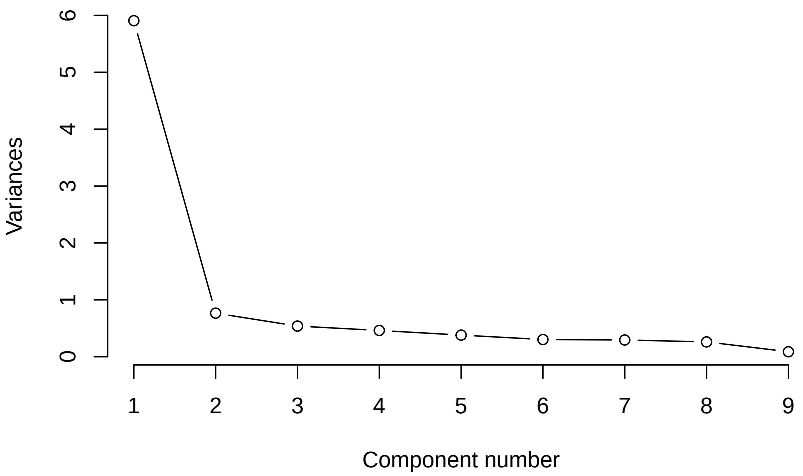
```
## [1] TRUE
```

Now that scaling is done perfectly, we can proceed with implementing different classification models and assess them.

Logistic Regression using PCA

Now that the data is scaled, we will implement Logistic regression model using PCA. Since, the generalized variance and total variance indicate that only few variables might contribute towards the majority of the variance in the data set, we will use PCA as a method of subset selection to pick variables that will help us the most in predicting whether a tissue sample is Benign or Malignant.

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4302 0.87512 0.73417 0.67979 0.61688 0.55010 0.54274
## Proportion of Variance 0.6562 0.08509 0.05989 0.05135 0.04228 0.03362 0.03273
## Cumulative Proportion  0.6562 0.74132 0.80121 0.85256 0.89484 0.92847 0.96120
##                            PC8     PC9
## Standard deviation     0.51074 0.29730
## Proportion of Variance 0.02898 0.00982
## Cumulative Proportion  0.99018 1.00000
```



The above summary and scree diagram of Principal components suggests that 3 principal components account for 80% variance in the data set. This means we can use 3 variables, instead of 9, and still preserve 80% of the variance in the data set. Let's have a closer look at these components to understand the relationship among the variables.

```
##                        PC1         PC2           PC3
## Cl.thickness    -0.3018371 -0.14815723  0.8652105309
## Cell.size       -0.3805230 -0.04953183 -0.0205078624
## Cell.shape      -0.3773482 -0.08449567  0.0327635671
## Marg.adhesion   -0.3325843 -0.04715193 -0.4127937472
## Epith.c.size    -0.3358808  0.15880282 -0.0864691991
## Bare.nuclei     -0.3350417 -0.25534370 -0.0009153633
## Bl.cromatin     -0.3456145 -0.22651135 -0.2153688808
## Normal.nucleoli -0.3353311  0.03131929 -0.1341740568
## Mitoses         -0.2326531  0.90748343  0.0874862451
```

The first principal component accounts for overall correlation among the variables since, all the values are quite close, except Mitoses. The Cell size and Cell shape have the largest values in the first principal component, however, other variables are not too far behind. The second principal component has a significantly large value for Mitoses, suggesting that cells with large Mitoses values are grouped together. Since the value of Mitoses is positive in the second principal component, it suggests an inverse relationship with all other variables in the same direction. The third principal component has large values for Clump thickness and Marginal adhesion, and thus accounting for variation in those 2 variables. Marginal adhesion has a negative value, indicating a direct relationship with the other variables in first two principal components. Clump thickness, on the other hand, has a positive value in the third principal component. This suggests an inverse relationship with other variables and, Clump thickness is significantly important in deciding whether a tumor is benign or malignant.

We will now transform our data set to fit on the new axes defined by the principal components selected.

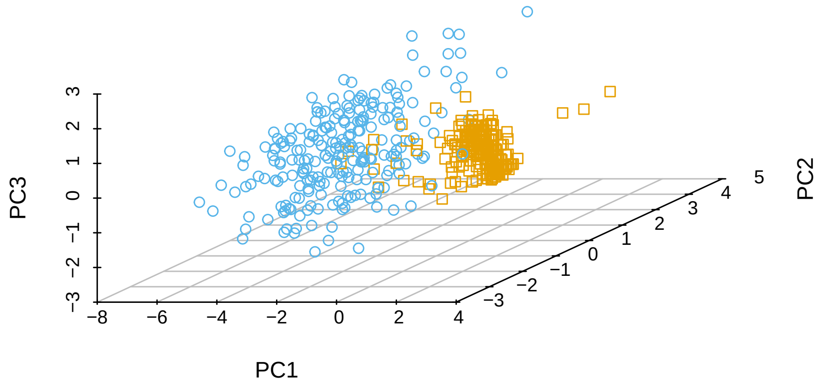
```
##         PC1         PC2         PC3 Class
## 1  1.470996 -0.11283640  0.56374313     0
## 2 -1.437451 -0.57522452 -0.24054863     0
## 3  1.593061 -0.07786382 -0.04996695     0
## 4 -1.474889 -0.55362197  0.59701026     0
## 5  1.345796 -0.09323342 -0.03119352     0
## 6 -5.004978 -1.55226476 -0.47430230     1
```

Finally, we will fit the logistic regression model on the data set.

```
log_reg_pca = glm(Class ~ ., data = BreastCancer_transformed,
                  family = "binomial")
```

```
##
## Call:
## glm(formula = Class ~ ., family = "binomial", data = BreastCancer_transformed)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1432     0.2846  -4.017  5.9e-05 ***
## PC1          -2.3118     0.2276 -10.159  < 2e-16 ***
## PC2          -0.3717     0.3746  -0.992   0.3212
## PC3           0.7276     0.3014   2.414   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 113.12  on 679  degrees of freedom
## AIC: 121.12
##
## Number of Fisher Scoring iterations: 8
```

The coefficients represent the log-odds of the response variable being in a particular category compared to the reference category. All the coefficients have a negative value, suggesting an increase in the predictor variable is associated with a decrease in the log-odds. The first principal component has the largest magnitude of the coefficient, suggesting strong correlation of Cell size, cell shape, marginal adhesion and Bare nuclei with the response variable. The third principal component has the second largest coefficient indicating strong correlation of Clump thickness and Marginal adhesion with the response variable. The p-value of first and third principal components are quite small, suggesting that Clump thickness and Marginal adhesion are very likely to be associated with the response variable.



The above scatter plot shows clustering of malignant tissue samples towards right side of the plot. Continuing with our interpretation of the PCs, we see that large values of Mitoses and Clump thickness are grouped together. Data points with overall negative values for principal component 1 are grouped towards the left of the scatter plot.

Logistic Regression using best subset selection

Logistic Regression with PCA showed quite decent results, however, it would be interesting to investigate what variables does best subset selection method choose and, whether we can see patterns similar to PCA. We can perform best subset selection of generalized linear models using **bestglm** package. We will use BIC (Bayesian Information Criterion) to select models with less number of parameters. BIC penalizes models with more parameters, aiming to balance goodness of fit and model complexity. Small values of BIC indicate better models.

```
## Morgan-Tatar search since family is non-gaussian.
```

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 0       TRUE        FALSE     FALSE      FALSE         FALSE        FALSE
## 1       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 2       TRUE        FALSE      TRUE      FALSE         FALSE        FALSE
## 3       TRUE         TRUE      TRUE      FALSE         FALSE        FALSE
## 4       TRUE         TRUE     FALSE       TRUE         FALSE        FALSE
## 5*      TRUE         TRUE     FALSE      FALSE          TRUE        FALSE
## 6       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 7       TRUE         TRUE     FALSE       TRUE          TRUE        FALSE
## 8       TRUE         TRUE     FALSE       TRUE          TRUE         TRUE
## 9       TRUE         TRUE      TRUE       TRUE          TRUE         TRUE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood      BIC
## 0        FALSE       FALSE           FALSE   FALSE    -442.17509 884.3502
## 1        FALSE       FALSE           FALSE   FALSE    -127.37980 261.2861
## 2         TRUE       FALSE           FALSE   FALSE     -83.15598 179.3649
## 3         TRUE       FALSE           FALSE   FALSE     -67.77778 155.1351
## 4         TRUE        TRUE           FALSE   FALSE     -61.37155 148.8491
## 5*        TRUE        TRUE            TRUE   FALSE     -56.13177 144.8960
## 6         TRUE        TRUE            TRUE   FALSE     -53.57186 146.3027
## 7         TRUE        TRUE            TRUE    TRUE     -51.63998 148.9654
## 8         TRUE        TRUE            TRUE    TRUE     -51.45031 155.1126
## 9         TRUE        TRUE            TRUE    TRUE     -51.44991 161.6383
```

The model with 5 predictors has the lowest BIC value, thus, we can extract the variables according to the best fitting model and construct a reduced data set.

```
## 
##  Predictors in the Best fit Model:
```

```
##    Intercept Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 5*      TRUE         TRUE     FALSE      FALSE          TRUE        FALSE
##    Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses logLikelihood     BIC
## 5*        TRUE        TRUE            TRUE   FALSE     -56.13177 144.896
```

```
## 
##  Indices of the Best fit Model:
```

```
## [1]  TRUE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
```

```
## 
##  Reduced Data Set:
```

```
##   Cl.thickness Marg.adhesion Bare.nuclei Bl.cromatin Normal.nucleoli
## 1    0.1977598   -0.63889730  -0.6983413   -0.181694      -0.6124785
## 2    0.1977598    0.75747664   1.7715689   -0.181694      -0.2848960
## 3   -0.5112687   -0.63889730  -0.4239068   -0.181694      -0.6124785
## 4    0.5522740   -0.63889730   0.1249621   -0.181694       1.3530163
## 5   -0.1567545    0.05928967  -0.6983413   -0.181694      -0.6124785
## 6    1.2613024    1.80475710   1.7715689    2.267589       1.3530163
##   y_bestsubset
## 1            0
## 2            0
## 3            0
## 4            0
## 5            0
## 6            1
```

The reduced data set contains only Clump thickness, Marginal Adhesion, Bare nuclei, Bland cromatin and Normal Nucleoli. Now, we will fit the logistic regression model using the reduced data set.

```
logreg_fit = glm(y_bestsubset ~ ., data = BreastCancer_red, family = "binomial")

summary(logreg_fit)
```

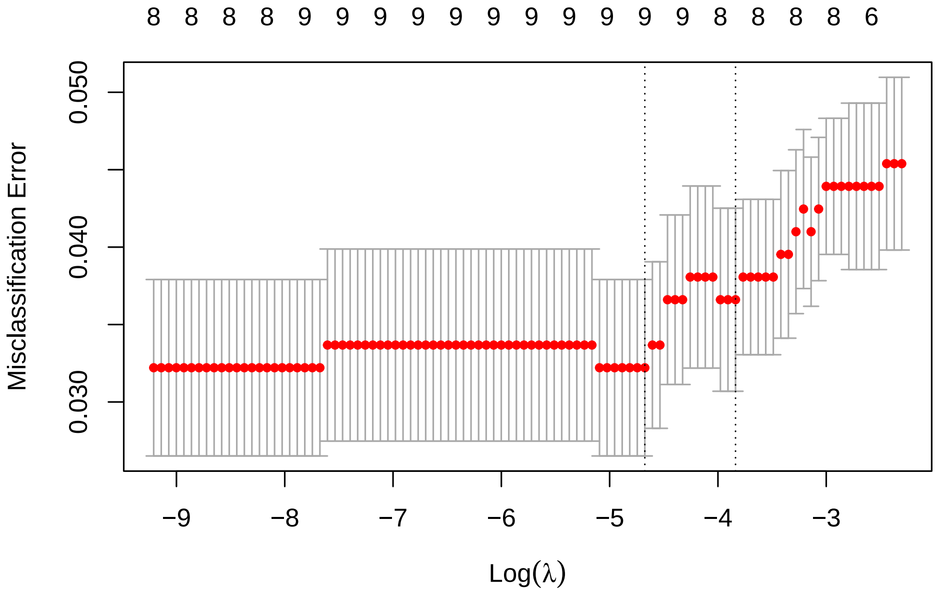
```
##
## Call:
## glm(formula = y_bestsubset ~ ., family = "binomial", data = BreastCancer_red)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.2700     0.2825  -4.495 6.97e-06 ***
## Cl.thickness      2.0910     0.3720   5.621 1.90e-08 ***
## Marg.adhesion     1.1319     0.3321   3.409 0.000652 ***
## Bare.nuclei       1.6300     0.3206   5.085 3.68e-07 ***
## Bl.cromatin       1.3544     0.3679   3.681 0.000232 ***
## Normal.nucleoli   1.0202     0.2986   3.417 0.000634 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 112.26  on 677  degrees of freedom
## AIC: 124.26
##
## Number of Fisher Scoring iterations: 8
```

All the 5 coefficients have a positive value, which indicates that an increase in any of these variables is associated with an increase in the log-odds of the cell being malignant. The coefficients have a significant difference in magnitude, with Clump thickness being the largest. All 5 coefficients have a very small p-value, however Marginal adhesion has the largest value of p among the other variables, suggesting that it is slightly less likely to be associated with the response variable compared to Clump thickness, Bare nuclei and cell size.

Logistic Regression with LASSO penalty

The next model that we are going to implement is Logistic Regression with LASSO penalty. LASSO penalty is a form of regularization technique which is used to improve the performance of the model. Regularization methods add a penalty to the loss function, in this case, it is the negative of the log-likelihood function. We will use LASSO method since we know that all the predictor variables don't contribute equally towards the overall variance in the data set and hence, LASSO regularization would perform variable selection and add shrinkage to the coefficients as well.

First, we will try to find the optimal value of lambda, which is the tuning parameter. We will also plot the Misclassification error to understand how the model behaves as the tuning parameter is increased.



We can see that the error reaches its maximum around value of -1 for Log(λ). Now, we will identify the optimal value for the tuning parameter and obtain the corresponding parameter estimates from the model.

```
##
##  Optimal Lambda value:
```

```
## [1] 0.009326033
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)     -1.0556726
## Cl.thickness     1.0727333
## Cell.size        0.2467263
## Cell.shape       0.7267145
## Marg.adhesion    0.4825691
## Epith.c.size     0.1596296
## Bare.nuclei      1.1564844
## Bl.cromatin      0.6922691
## Normal.nucleoli  0.4553087
## Mitoses          0.1755041
```

For the optimal value of lambda, we find that LASSO penalty has reduced the values of coefficients significantly. Clump thickness and Bare nuclei have the largest values of coefficients, meaning they contribute significantly towards prediction of a tissue sample being benign or malignant. Cell size, Epithelial cell size and Mitoses have the lowest coefficient values and won't contribute much towards the prediction. All the coefficients are positive indicating a positive relationship with the target variable. Cell shape and Bland cromatin also have significant coefficient values. These findings are in sync with our previous models using subset selection and PCA.

Quadratic Discriminant Analysis

Next, we are going to implement Discriminant Analysis. Discriminant Analysis is a classification technique that aims to find the optimal linear combinations of features that best separate different classes in a data set. To decide whether to use Linear Discriminant Analysis or Quadratic Discriminant Analysis, we will have a look at the covariance matrix of each class.

```
##
##  Comparison of Covariance Matrix of Benign and Malignant Class:
```

```
## [1] "Mean relative difference: 5.731006"
```

LDA makes the assumption that the covariance matrices of each class is equal. Since the covariance matrix of each class is not equal, as displayed above, it would be better to use QDA rather than LDA. Also, since we know from previous subset selection methods that we do not need to include all the variables in the model, we will select the variables we extracted when we performed best subset selection method.

```
## Call:
## qda(y_bestsubset ~ ., data = BreastCancer_red)
##
## Prior probabilities of groups:
##         0         1
## 0.6500732 0.3499268
##
## Group means:
##   Cl.thickness Marg.adhesion Bare.nuclei Bl.cromatin Normal.nucleoli
## 0   -0.5240440    -0.5178153  -0.6031546   -0.555890      -0.5268939
## 1    0.9735377     0.9619665   1.1205047    1.032699       0.9788322
```

The prior probabilities of groups represent the estimated probability of each class occurring before observing any data. Prior Probability of Benign class is 65% and that of Malignant class is 35%. Next most important values are the Group means. Group means represent the estimated mean values of each predictor variable for each class. These values give you a sense of the central tendency of each class with respect to the predictor variables. The Group means of Clump thickness and Marginal adhesion are quite near each other. On the other hand, Bare nuclei has slightly higher values for group means. Normal Nucleoli and Bland cromatin have group means roughly in the same range as Clump thickness and Marginal adhesion.

Performance evaluation using Cross Validation

To assess the performance of all the models created so far, we will perform Cross validation. To make the assessment fair, its important that all the models are trained and tested on the same subsets of the original data set. We will calculate the weighted average of the test errors and compare the models based on this value.

```
##
##  Logistic Regression using PCA :  0.03513909
```

```
##
##  Logistic Regression using Best Subset Selection :  0.03660322
```

```
##
##  Logistic Regression using LASSO :  0.03367496
```

```
##
```

```
##  Classification using QDA:  0.04685212
```

As per the test errors shown above, Logistic Regression with LASSO penalty has the smallest test error and thus would be the best choice for predicting whether a tissue sample is Benign or Malignant based on the cytological characteristics. Even though LASSO penalty did not reduce the coefficients to 0, it still performed better than all the other models by adding significant penalties on the coefficients of less important variables, thus shrinking their values to near 0. The LASSO model seems to generalize well and not overfit the data by removing variables, instead it has reduced the values of the coefficients of low impacting variables. Quadratic Discriminant Analysis has the largest average test error and performed the worst. One reason why QDA performed worst could be outliers. QDA is quite sensitive to outliers and as explored previously, our data set did contain a lot of outliers.

Conclusion

We have successfully explored the data and implemented various classification algorithms. We started by implementing PCA and best subset selection techniques to try and reduce the number of predictor variables. We also tried to add regularization to the Logistic Regression by adding LASSO penalty, however it did not result in removal of any predictor variables, just shrinkage of less important coefficients. Finally, we implemented QDA with best subset selection technique. To assess the performance of all these models, we tried cross validation based on test errors. Logistic Regression with LASSO penalty has the smallest test error, and a close second is Logistic Regression using PCA. We concluded that Logistic Regression with LASSO penalty is the best classifier, even though it has all the predictor variables.